

# Big Mirrors, Bayesian Evangelists and the Public:

How Advances in Mirror Technology, detectors,  
databases, machine learning and  
crowd-sourcing is driving Astronomy into the future

Michael Way (NASA/Goddard Institute for Space Studies)

<http://www.giss.nasa.gov/staff/mway/bbp2011.pdf>

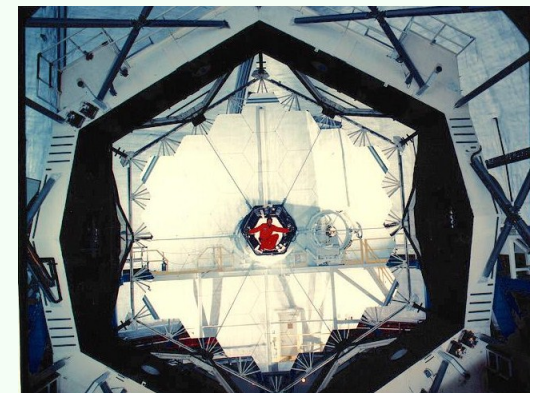
# Outline

- Mirrors: Single dish to segments
- Detectors: paving with plates or silicon?
- Data: From ascii files to databases
- Crowd-sourcing: AAVSO → Zooniverse
- Machine Learning
- Examples...

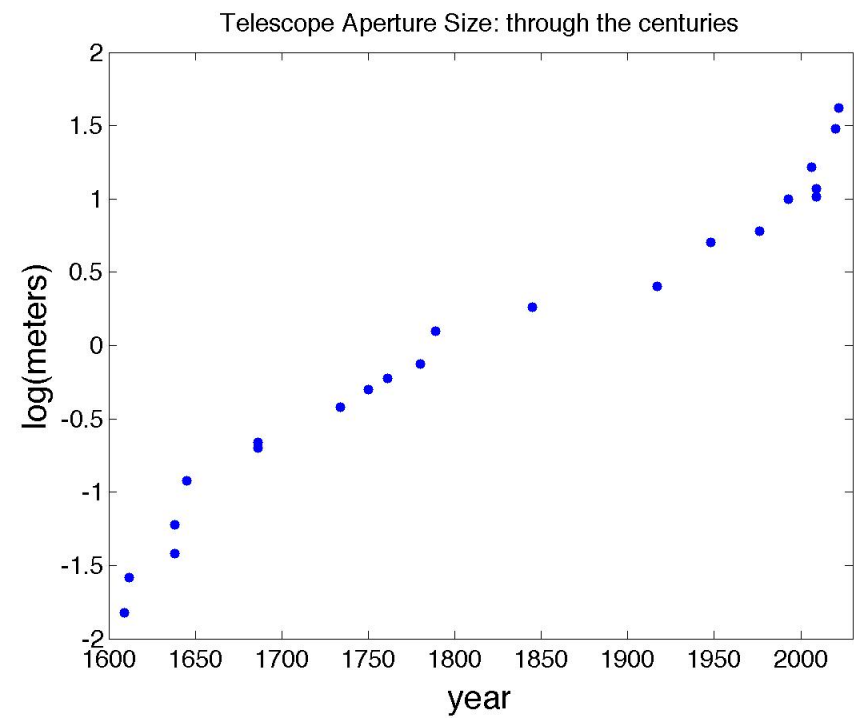
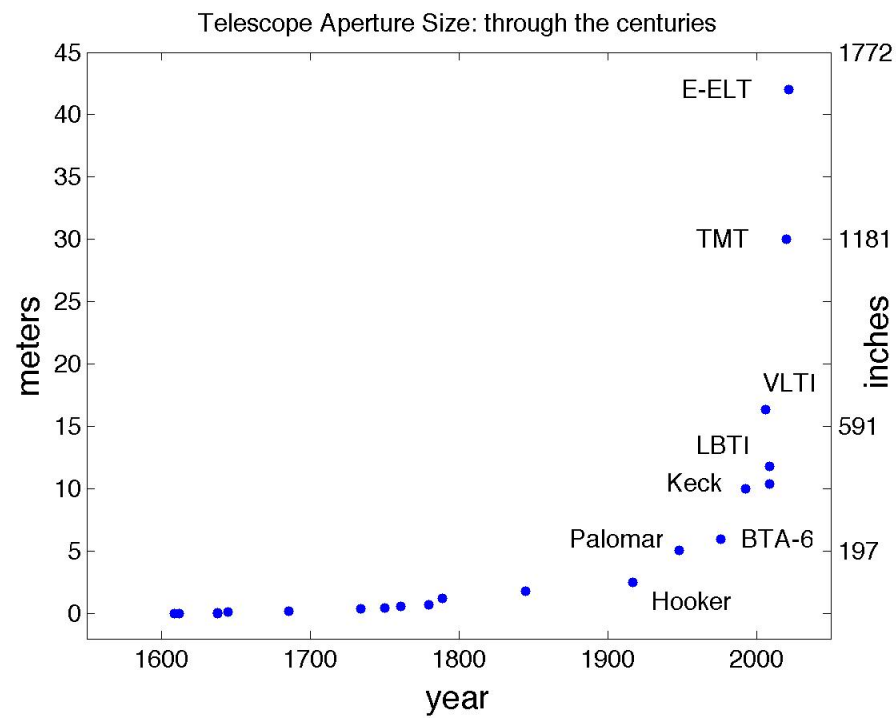
# Mirrors: single to segments



- Lippershey/Galileo refractor: 1.5cm? 1608/1609
- Newton/Hooke reflector: 3.3cm/18cm: 1668/1674
- Herschel: 1.26m: 1789
- Leviathan/Hooker: 1.83m/2.54m: 1845/1917
- Hale/BTA: 5.08m/6m: 1948/1976
- Keck/GTC: 10m/10.4m: 1993/2009
- TMT/E-ELT: 30m/42m: 2020?



# Technology





# Telescopes/Culture

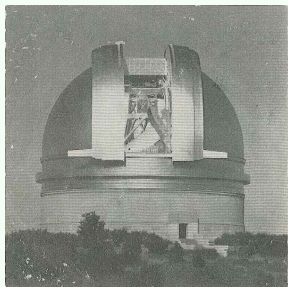
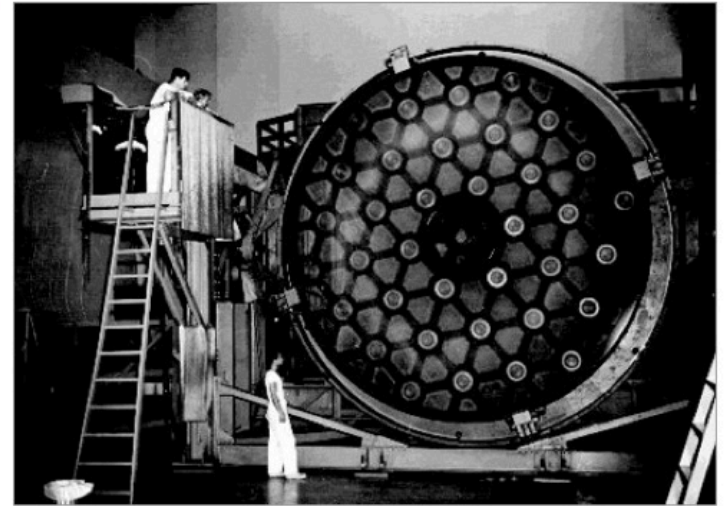


## Oct. 3, 1947: Birth of Palomar's 'Giant Eye'

By Tony Long October 3, 2011 | 6:30 am | Categories: 20th century, Astronomy, Engineering

**1947:** After 13 years of grinding and polishing, the Palomar Observatory mirror is completed at Caltech.

It was, at the time, the largest telescope mirror ever made in the United States, measuring 200 inches in diameter. Following its completion, the disk was mounted in Palomar's [Hale Telescope](#) and first used in January 1949 to take pictures of the Milky Way. [Edwin Hubble](#) was the first astronomer to make images using the new scope.

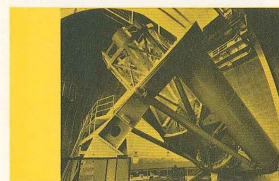


### The Hale Telescope

The telescope of which the 200-inch mirror is the heart was planned in 1929 by the veteran astronomer, George E. Hale. Funds were granted by the Rockefeller Foundation and Mt. Palomar in Southern California selected as the observatory site. At dedication ceremonies on June 3, 1948, the telescope was named in memory of Dr. Hale who died ten years before completion of the project.

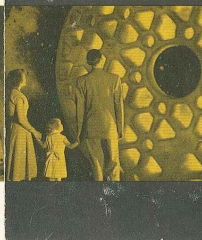
Galileo's primitive telescope lens measured only 2 3/4 inches in diameter (smaller than this souvenir disk) and could see 81 times that of the human eye. Recent observations indicate that the Hale Telescope is picking up stars 6 million times dimmer than the naked eye can spot.

Using photographic plates, astronomers can see star groups four times as faint and twice as far away as was previously possible.



CORNING GLASS CENTER • Corning, N. Y.  
Open 9:30-5:00 daily except Mondays

### The 200-Inch Disk



in  
the  
CORNING  
GLASS  
CENTER

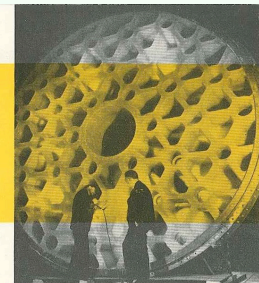


### The 200-Inch Disk

The giant glass disk in the lobby of the Corning Glass Center, Corning, N. Y., is the first casting of a telescope mirror produced for the California Institute of Technology. Measuring 17 feet in diameter and 26 inches thick, the 20-ton disk is the largest piece of glass made by man.

Corning Glass Works was commissioned in 1931 to manufacture the mirror of Pyrex brand glass (similar to the type in your baking and dinner ware) because it holds its shape regardless of temperature changes and lends itself to precision polishing.

The transparent characteristics of glass were not considered, as the telescope was to be the reflecting type in which the mirror would concentrate light rays striking its concave surface. To reduce its weight and shorten cooling time after casting, a "waffle iron" ribbed design was developed for the rear of the mirror.

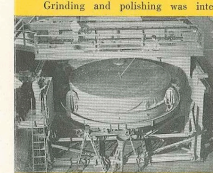


Experimental disks of 30-, 60- and 120-inch diameters were first produced and finally on March 25, 1934, white hot molten glass was poured into the 200-inch mold of insulated brick. The work was nearly completed when the intense heat melted several steel bolts anchoring the mold, and pieces of the core floated to the surface. (This accounts for the solid spots in the ribbed surface of the Glass Center disk.)

An improved mold was designed and the second disk was cast successfully on December 2, 1934. The huge piece of glass was held at 1200 degrees Fahrenheit for 60 days and cooled for eight months by dropping the temperature only slightly more than one degree each day.

The rough mirror was shipped from Corning in a special "well" car on March 12, 1936 and arrived at Pasadena two weeks later where it was installed in the California Institute of Technology optical laboratory. Then began the pains-

taking work of hollowing the mirror face to the proper curvature. Grinding and polishing was inter-



rupted by the war and completed in October, 1947 after 5 1/2 tons of glass had been ground away.

# Detectors

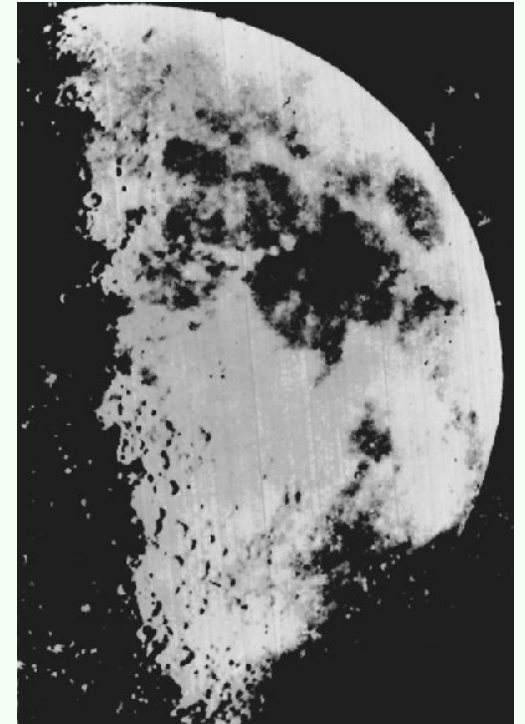
- Eye
- 1840: Photographic Plates
- 1890--1980: Photoelectric Photometers
  - Reticons too!
- 1970: CCDs



FIG. 13.—RCA 1P21 photomultiplier (see Engstrom 1947)

# Detectors/QE: Paving the focal plane

- Eye: QE 1—10% : size: 0.7cm
- Glass Plates (Wet) QE 1%
  - 1839 Daguerre: Moon images
  - 1840 Draper
    - 13cm reflecting telescope
    - Moon (20 min)
  - 1850: Bond and Whipple: first star photo - Vega
  - 1872 Miller/Huggins: First spectrograph – Sirius



Earliest known Daguerre image: 1851 by John Adam Whipple -----↑

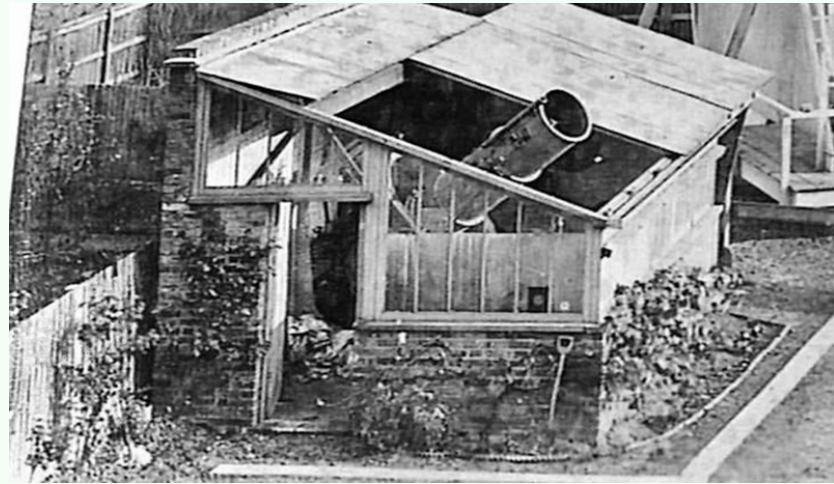


# Detectors/QE: Paving the focal plane

- Glass Plates (Dry) QE 1—3%
  - 1876 Huggins (spectrograph)
  - 1883 Common : First objects fainter than seen by eye
    - 91cm reflecting/60 min exposure



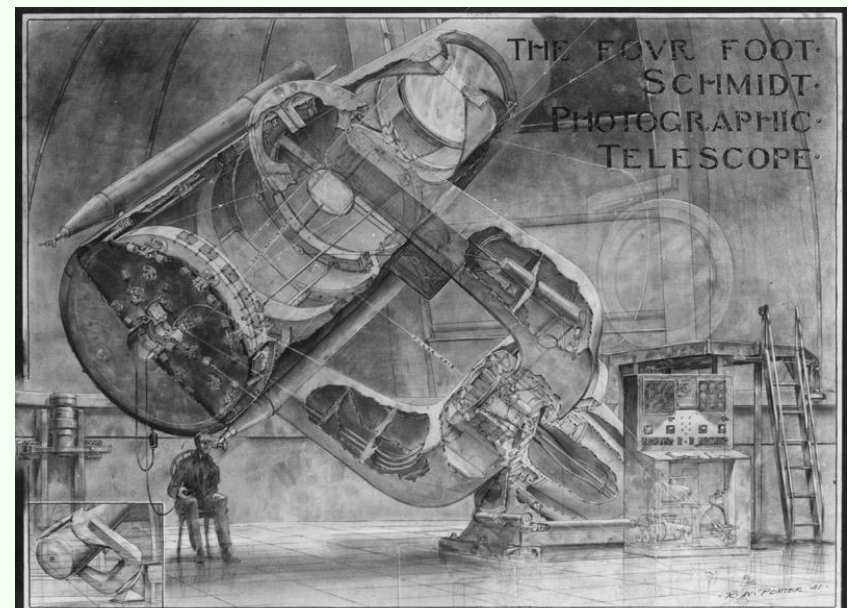
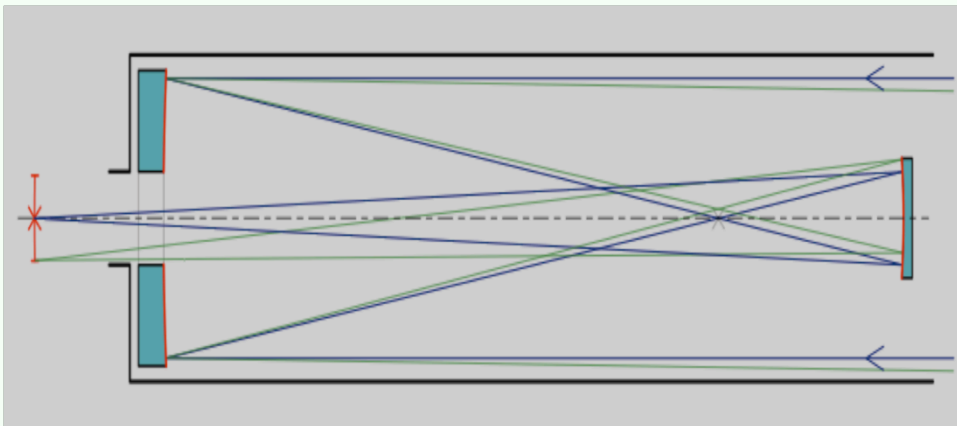
Orion Nebula by Common



Early Common Observatory

# Detectors/QE: Paving the focal plane

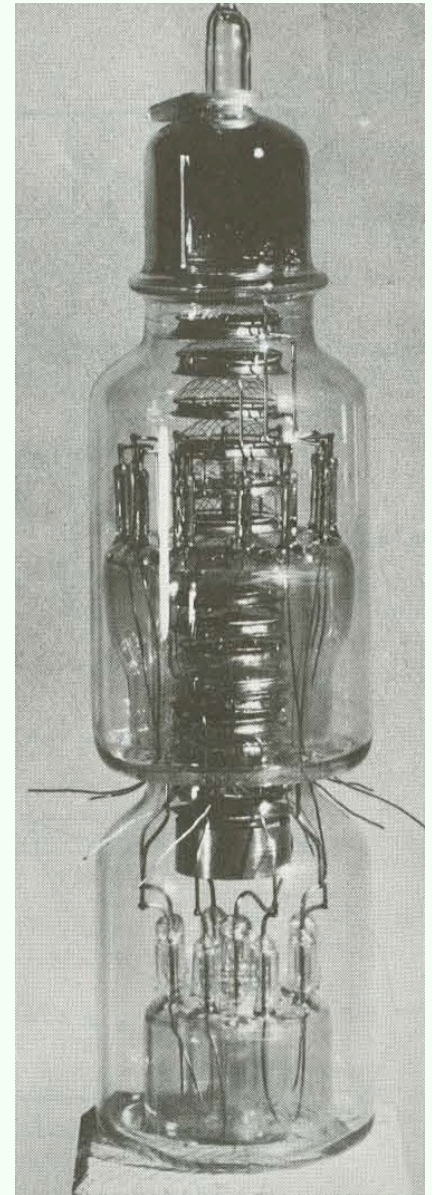
- Glass Plates (Dry) QE 1—3%
  - 1887 Astrographic Catalogue and Carte du Ciel
    - Aperture  $\sim 33$  cm, scale: 60 arcsec/mm
    - Field of view:  $2^\circ \times 2^\circ$  ! (Moon is nearly  $1.5^\circ \times 1.5^\circ$ )
  - 1948 Oschin Schmidt (Palomar): 1.22m, FoV=  $4^\circ \times 4^\circ$  degrees
    - 14" x 14" glass plates (41"/mm) – paving the focal plane with glass
  - 1970 DuPont 2.54m ( $2.1^\circ \times 2.1^\circ$ )



# Detectors

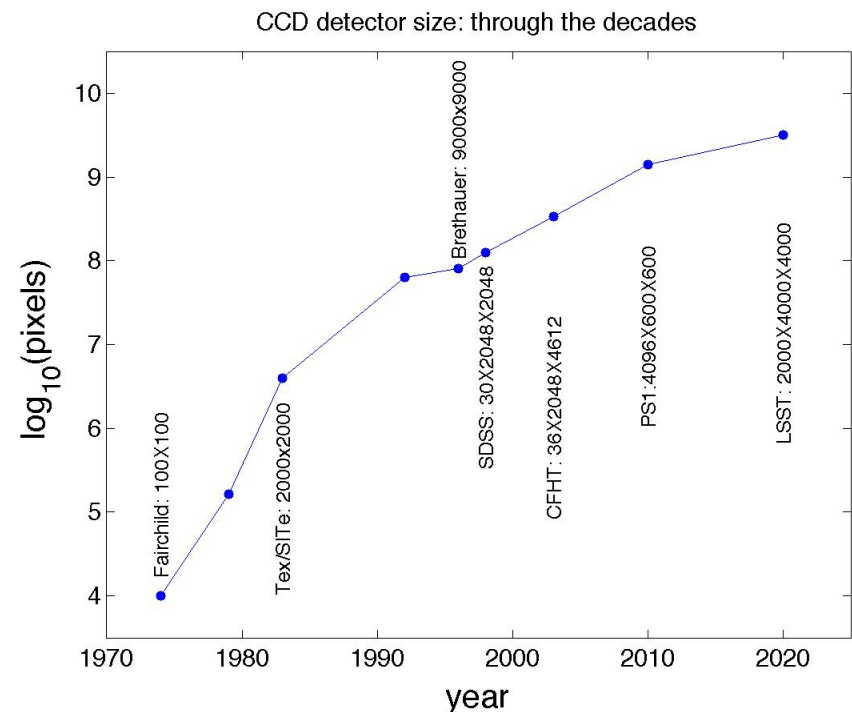
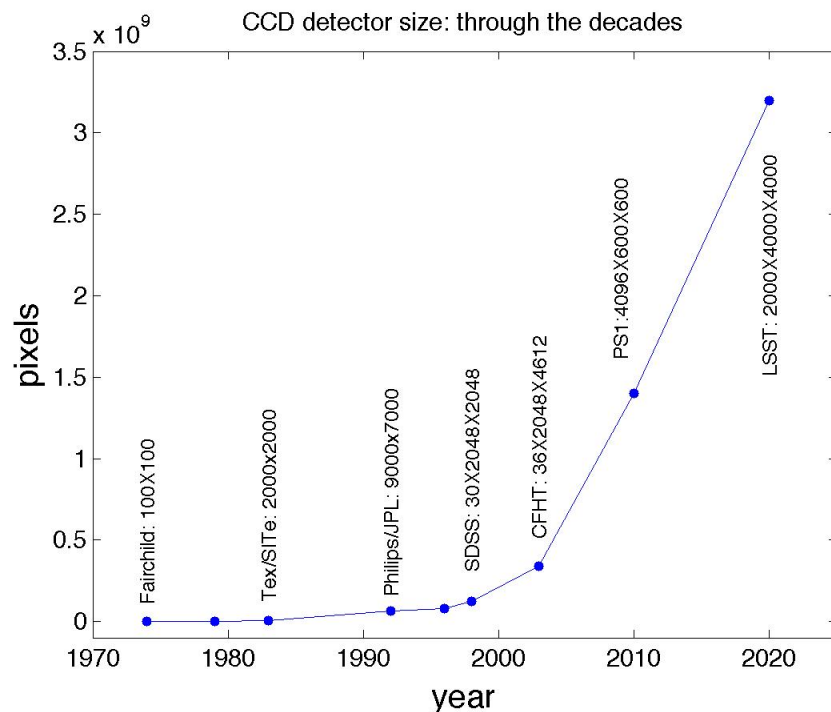
- 1892 – 1980s: Photoelectric photometers
  - Think about it as a single element CCD
  - Digital output? More likely paper tape...
  - X-rays: V2 rockets used in 1949!

19 stage linear photomultiplier tube developed at the Paris Obs ➔



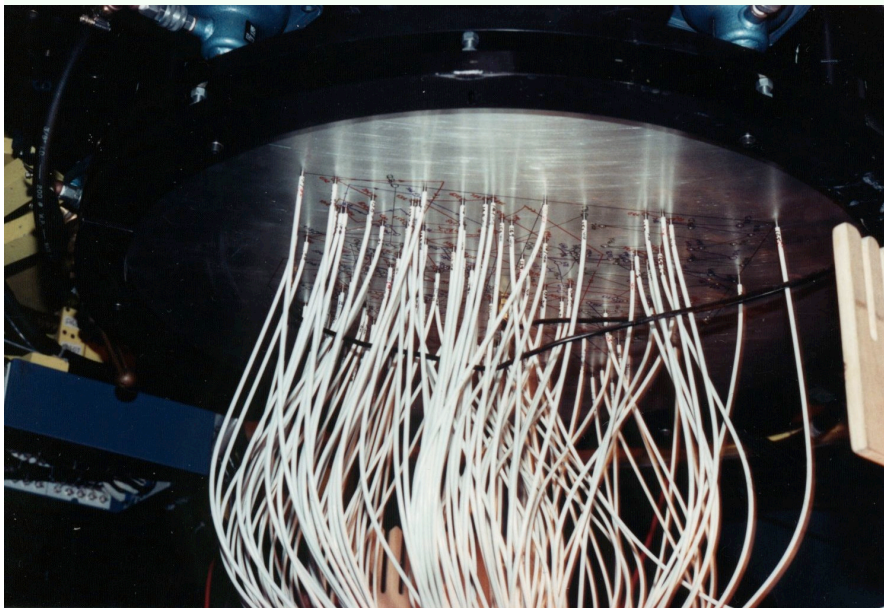
# Detectors/QE: Paving the focal plane

- Charged Coupled Devices (CCDs) QE: 10—40%
  - Boyle and Smith 2009 Nobel Prize for 1970 CCD development at Bell Labs (the 7<sup>th</sup> from this lab)
- From Small beginnings to filling the focal plane:

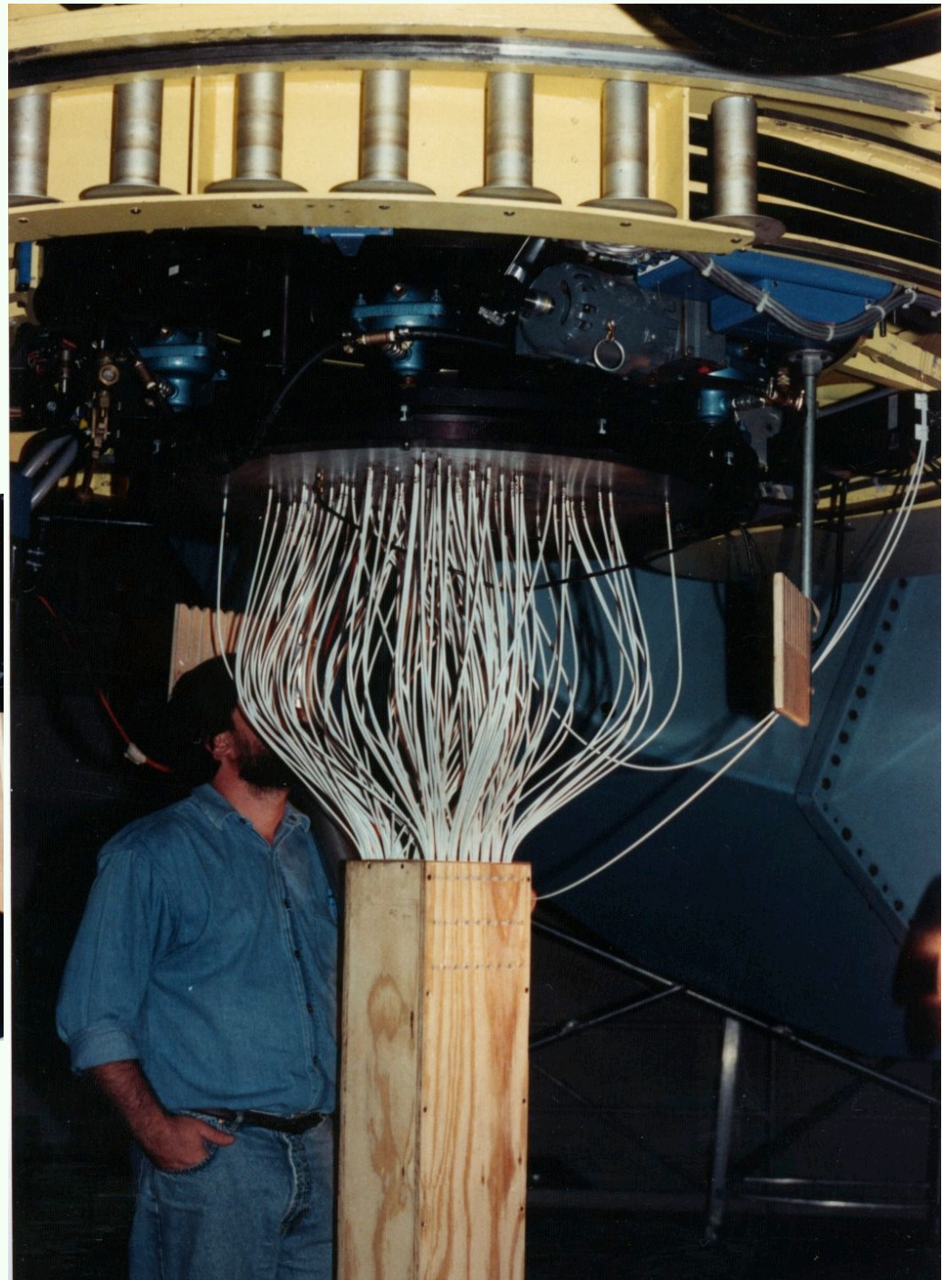




Before CCDs were big enough we filled the focal plane in other ways: e.g. photographic glass plates and Fiber Spectrographs:



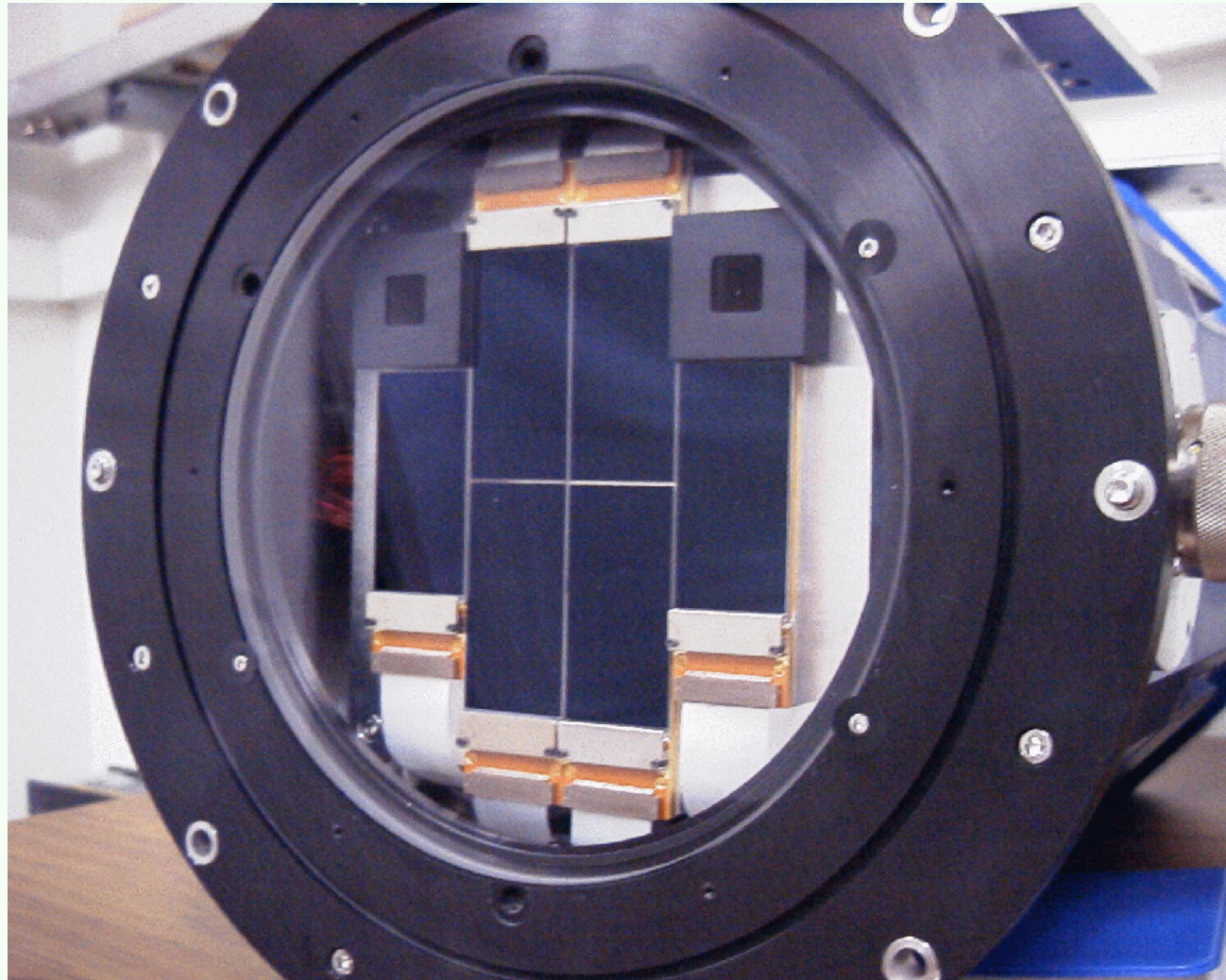
Dupont 2.54 meter  
Las Campanas, Chile





# Hale 5 meter CCD detectors

Today 6 x SITE 2048x4096 CCDs fills much of the Hale 5 meter Field of View: But it is only ~12cm x 12cm



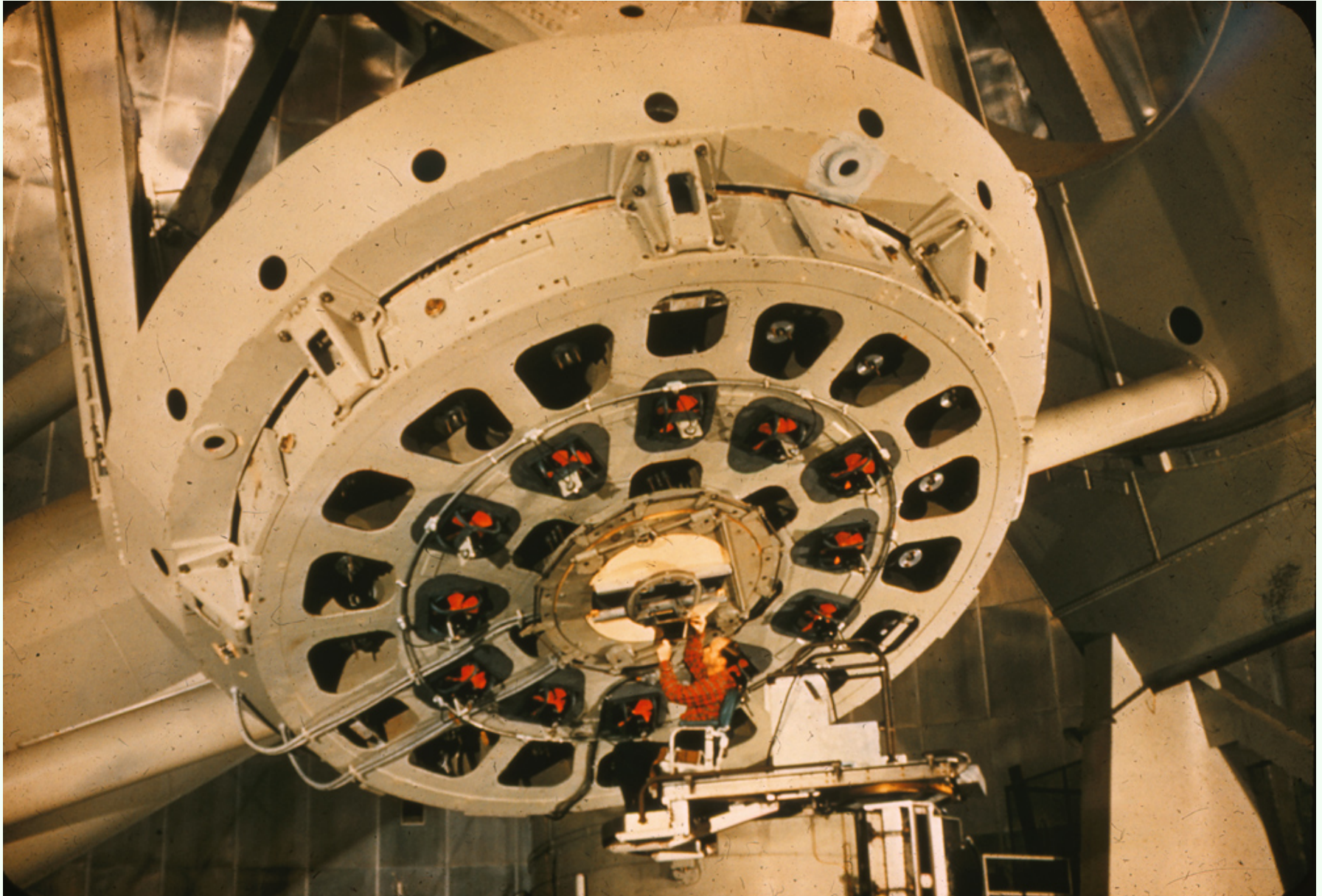


# Hale 5 meter photographic plate





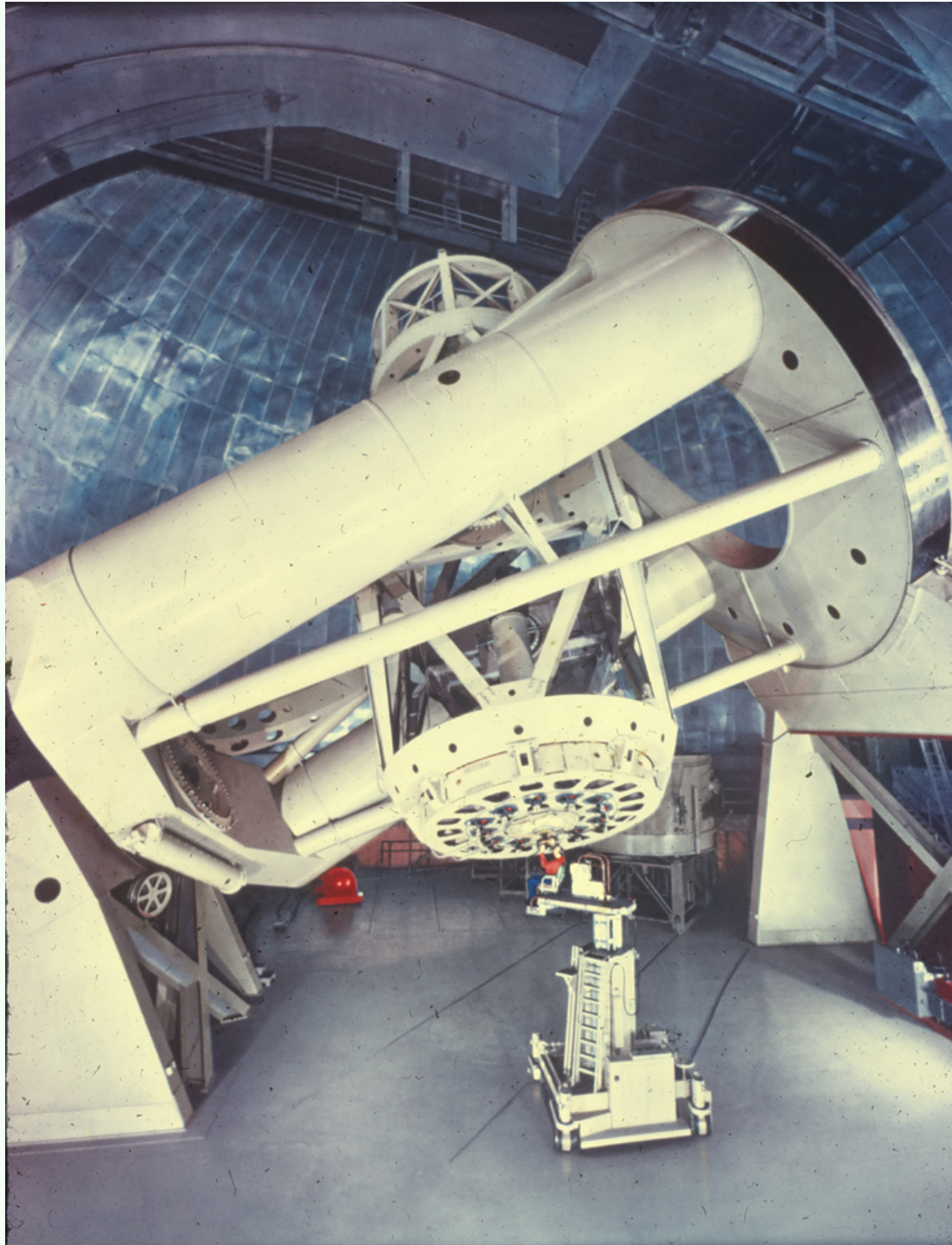
# Hale 5 meter photographic plate



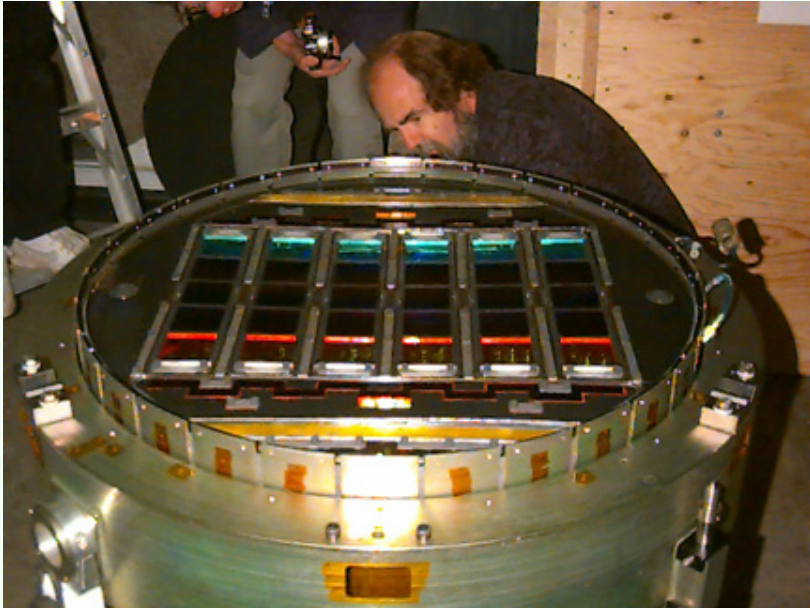


Det

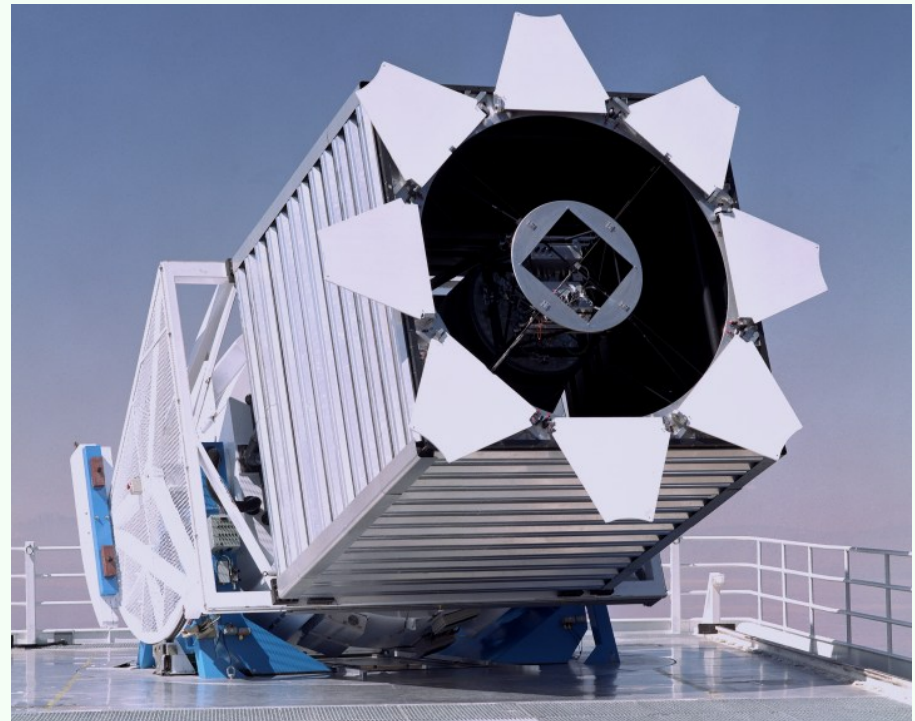
al plane



# Sloan Digital Sky Survey Telescope



← 30 2048x2048 CCDs



← 600 fiber spectrograph

Data:  
From ascii files to databases  
(Revolutions in media storage)

- CCDs (and some photomultipliers) gave us a fully electronic record and portable storage was catching up too!
  - 9 track tapes [1970] (800-6250 bpi) = 170MB
  - 4mm DAT = 90m (2GB), 120m (4GB with DAT2)
  - 8mm Exabyte = 112m (2.5/5GB), 54m (1.2/2.4GB)
  - CD = 700MB (100 year lifespan?)
  - DVD = 4.7GB
  - Today? It all sits on disk (**or in the cloud**)...



# Revolutions in Data Handling

- Traditionally: all data acquisition and reduction was handled by Astronomers on their own
- Today: most astronomical data is collected in surveys or by service observations



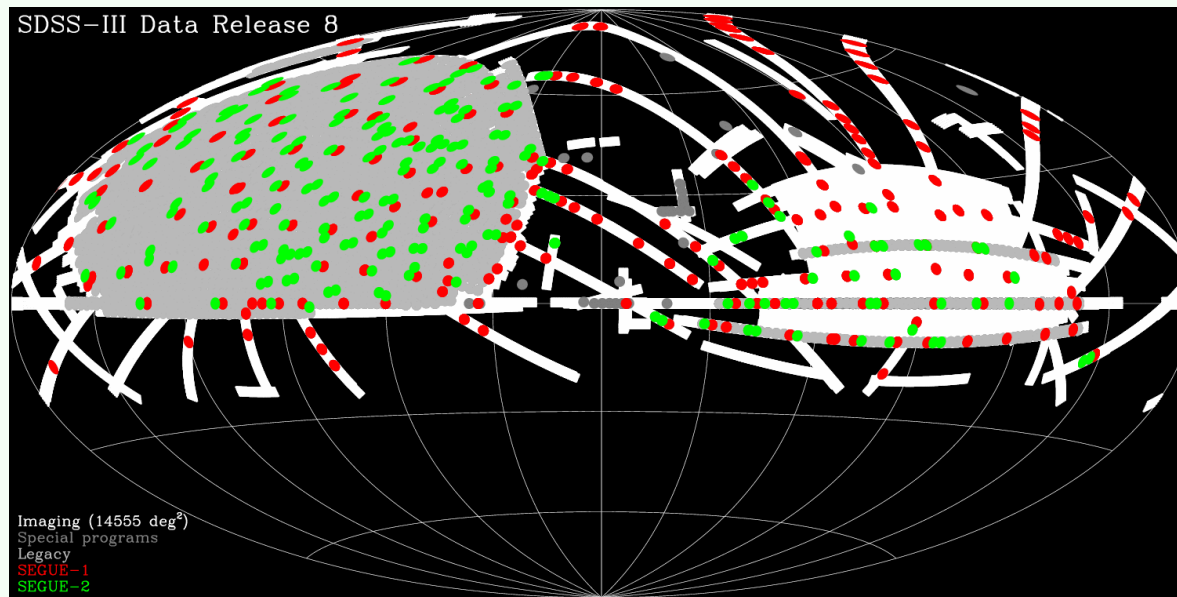
- Digitized Sky Survey 1980s: **102 CDROMs**
- APM Survey 1990s (Scans of Schmidt plates)
  - 150 million objects detected
  - Basis for the Two degree Field Galaxy Redshift Survey (2000)  
250,000 spectra
- Two Micron All Sky Survey (early 1990s):
  - first large scale fully digital survey (two 1.3m) and catalog.
  - 471 million objects detected
  - Released on **5 double-sided DVDs (43GB)**
  - Full fidelity images ~10TB

(dB too: arXiv:1110.4206v1)



# Sloan Digital Sky Survey (2000—Present)

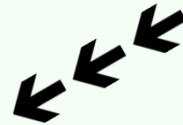
- 120 Megapixel camera, 1.5x1.5 degrees
- 600 multi-object spectrograph
- Data Release 7 (~30TB total, 3TB database)
  - 357 million unique objects, 1.6 million spectra
- Data Release 8: 930 million unique objects



1970-80s



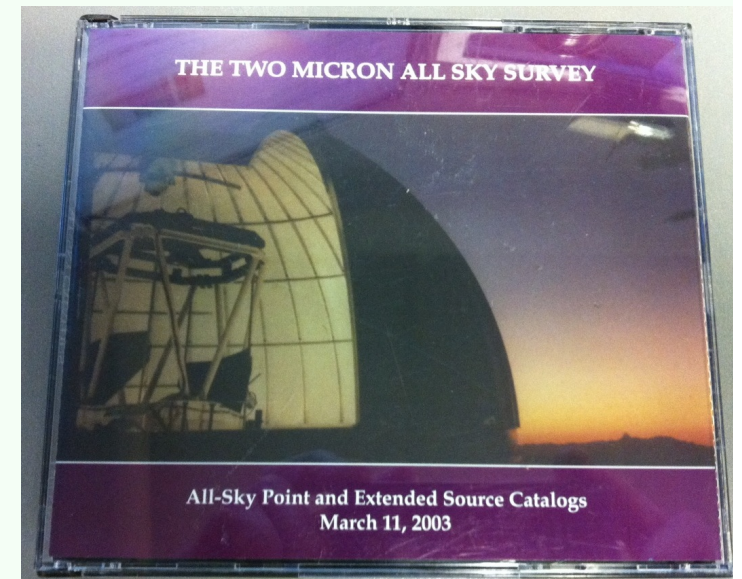
Late 1980s



1990s




2000s





# 2005 and beyond

SDSS Query / CasJobs

HelpToolsQueryHistoryMyDBImportGroupsOutputProfileQueuesSkyServerLogout

## 'qso1' Details

Resubmit Job

JobID	TaskName	Context	Queue	Submitted	Started	Finished	Status
5494773	qso1	DR7	600	5/20/2011 3:26:35 PM	5/20/2011 3:26:38 PM	5/20/2011 3:29:40 PM	Finished

Executed on	Rows	Message
DR7Best long	76454	Query Complete

Query

```
Select p.ObjID, p.ra, p.dec,
p.dered_u, p.dered_g, p.dered_r, p.dered_i, p.dered_z, p.petroR50_i, p.petroR90_i,
p.Err_u, p.Err_g, p.Err_r, p.Err_i, p.Err_z, p.petroR50Err_i, p.petroR90Err_i,
s.z, s.zErr, s.zConf, s.zStatus into mydb.qsol from SpecOBJall s, PhotoObjall p
WHERE s.specobjid=p.specobjid
and s.zWarning=0
and (SpecClass=dbo.fSpecClass('QSO') or SpecClass=dbo.fSpecClass('HIZ_QSO'))
and ((flags & 0x8) = 0) and ((flags & 0x2) = 0) and ((flags & 0x40000) = 0)
```

# Unexpected Collaborators too?

- SDSS dB was built in collaboration with Microsoft
  - Jim Gray with Alex Szalay and others...
  - Interesting large problem, open source data model
  - Still possible to download O(100s GB) data sets
- SciDB: built for next generation data sets (LSST)
  - <http://www.scidb.org>
  - Not possible to download PB sized data and use it
    - 1PB over 10Gb/s line is 10 days, 1PB = \$200 in 2020?
    - I/O not keeping up with other Moore type laws ([arXiv:1108.5124v1](https://arxiv.org/abs/1108.5124v1))
  - R-interface for expert users?!
  - **SciDB Community Meeting Oct 18<sup>th</sup> (SLAC)**

# Tomorrow...

- PanStarrs (2011—2020?)
  - 64x64 array (600x600 CCD) 1.4 Gigapixels
  - ~3TB/night
- Dark Energy Survey (2012—2017)
  - 74 CCDs
  - 1TB/night raw data
- Large Synoptic Survey Telescope (2020—2030)
  - 1PB/night raw data
  - Database size: ~10PB
  - 60PB of images



# Crowdsourcing: From AAVSO to Zooniverse

## Crowdsourcing is older than you think

- AAVSO = American Assoc. of Variable Star Observers
  - **Amateur Astronomers contributing observations of variable stars since 1911!**
- 1999: SETI @ Home and copycats
  - Called “Volunteer Computing”
  - Really just distributed computing, not crowdsourcing



# Modern Citizen Science

- 2000: Clickworkers (NASA/Ames DDF!)
  - Identifying & Classifying age of Martian craters from Viking Orbiter images
  - Kanefsky, Barlow and Gulick.
- 2006: Stardust@Home
  - Search aerogel images for tiny dust impacts gathered from tail of Comet Wild

- 2007: GalaxyZoo
  - Classifying Galaxies in the Sloan Digital Sky Survey
  - Largest by eye “professional catalog” ~1400 objects
  - To-date they have classified over 1 million objects
  - Use Machine Learning to train automated classifiers...
- Zooniverse does a lot more:
  - Planet Hunting
  - Transcribe old Weather Logs
  - Ancient lives (reading old papyri)



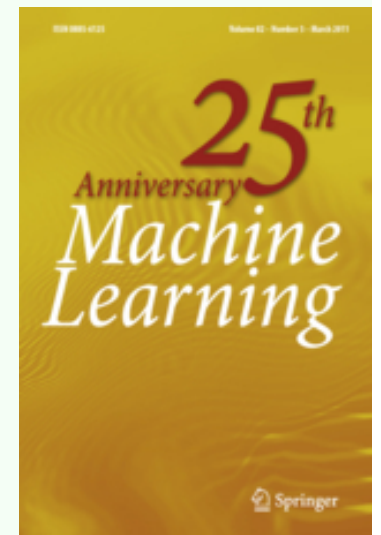
Rev. Bayes 1702-61

Least Squares



1795 (1809): Gauss  
1805: Legendre  
1808: Adrian

Shannon (1916-2001)



Machine Learning

Machine Learning

# Machine Learning

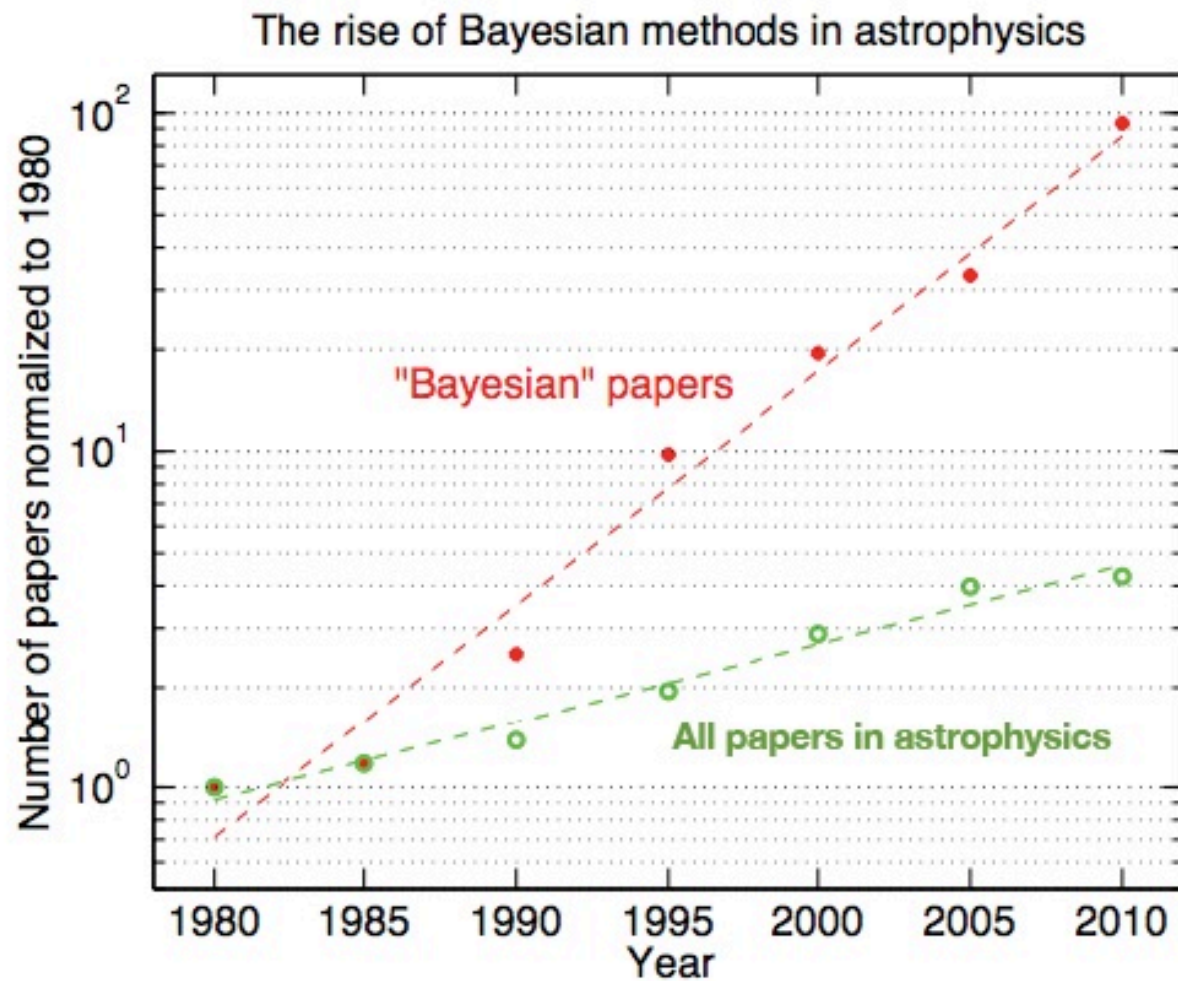
Given new large complex multivariate data  
Machine Learning & Data Mining are becoming  
more commonly used

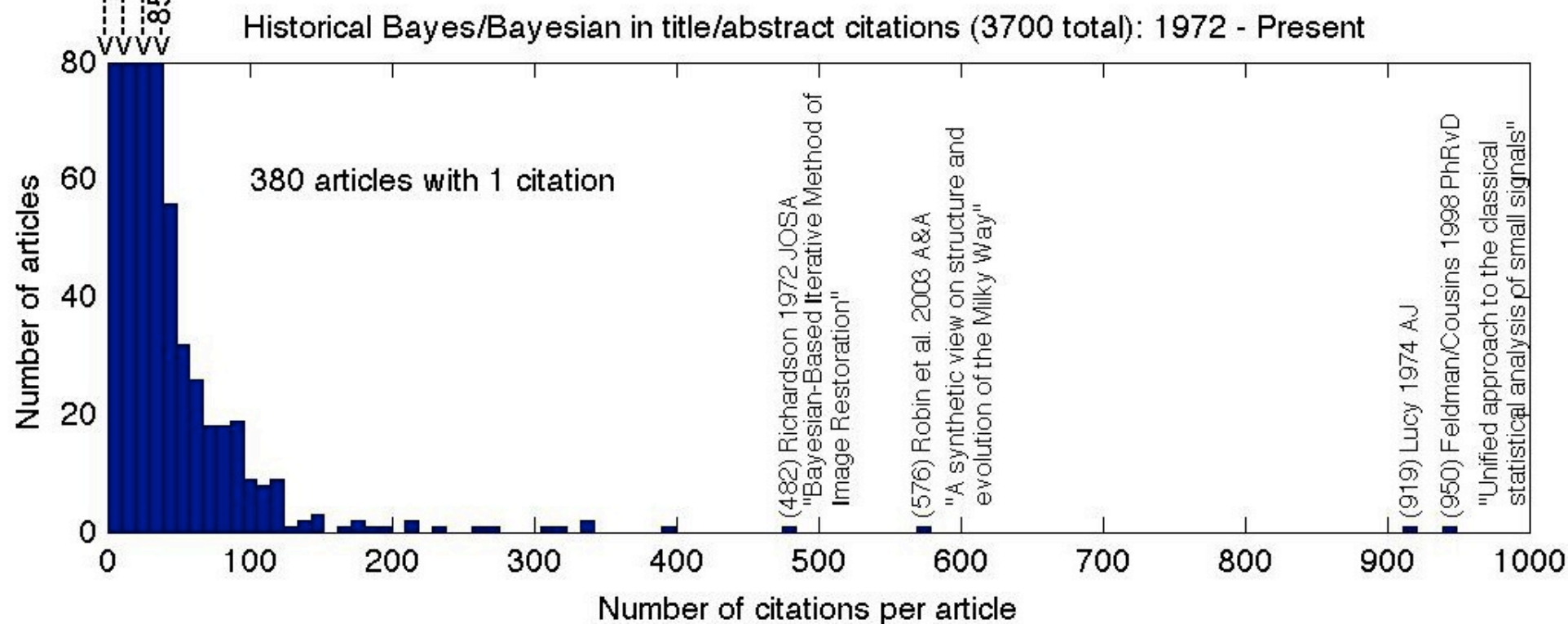
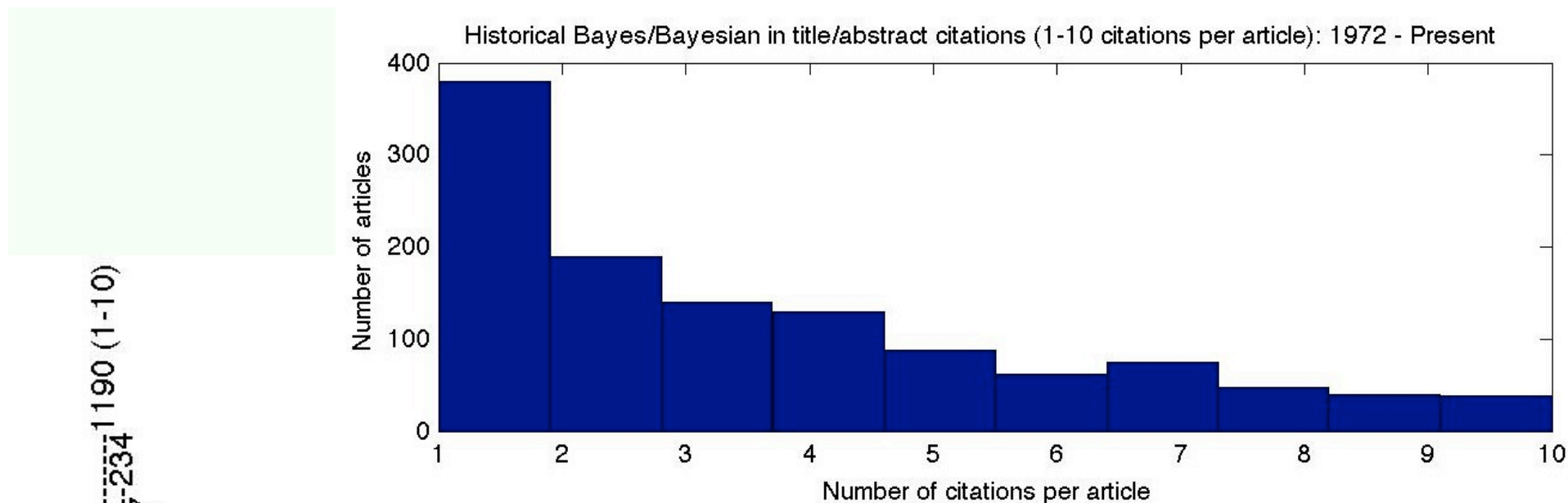


- 3700+ plus papers with Bayes in title or abstract
- Large numbers of citations for most popular papers
- Increase in number year by year...

# Reverend Bayes

Review of Bayesian methods in cosmology: Trotta (2008), arxiv: 0803.4089







# Lucy and Hubble

## **An iterative technique for the rectification of observed distributions**

L. B. Lucy\*

*Departments of Physics and Astronomy, The University of Pittsburgh, Pittsburgh, Pennsylvania 15213*

(Received 15 January 1974; revised 26 March 1974)

An iterative technique is described for generating estimates to the solutions of rectification and deconvolution problems in statistical astronomy. The technique, which derives from Bayes' theorem on conditional probabilities, conserves the constraints on frequency distributions (i.e., normalization and non-negativeness) and, at each iteration, increases the likelihood of the observed sample. The behavior of the technique is explored by applying it to problems whose solutions are known in the limit of infinite sample size, and excellent results are obtained after a few iterations. The astronomical use of the technique is illustrated by applying it to the problem of rectifying distributions of  $v \sin i$  for aspect effect; calculations are also reported illustrating the technique's possible use for correcting radio-astronomical observations for beam-smoothing. Application to the problem of obtaining unbiased, smoothed histograms is also suggested.

# Machine Learning

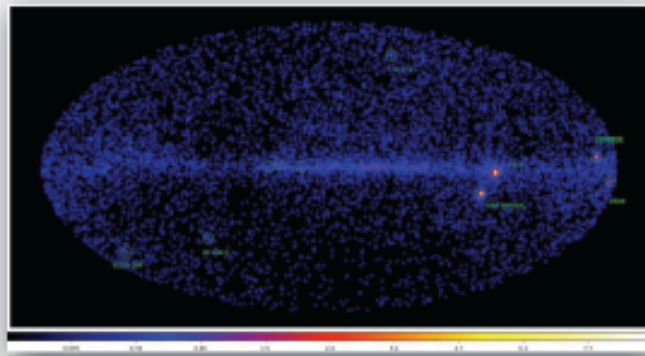
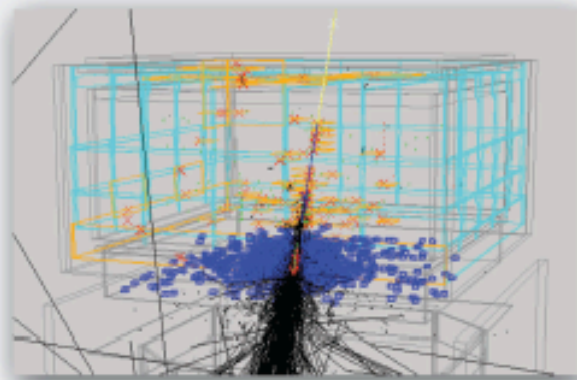
- 1395 with “Neural Networks” (well known?)
  - 9563 citations
- 1044 with “Data Mining”: 4709 citations
- 288 with “Machine Learning”: 1520 citations
- 108 papers with Self Organizing Maps (obscure?)
  - 457 citations

-----

- Statistical Challenges in Modern Astronomy
  - 5 conferences (every few years) 1991—2011 (+ book)
  - 7 astrostatistics summer schools



# Advances in Machine Learning and Data Mining for Astronomy



Edited by  
**Michael J. Way, Jeffrey D. Scargle,  
Kamal M. Ali, and Ashok N. Srivastava**

 **CRC Press**  
Taylor & Francis Group  
A CHAPMAN & HALL BOOK

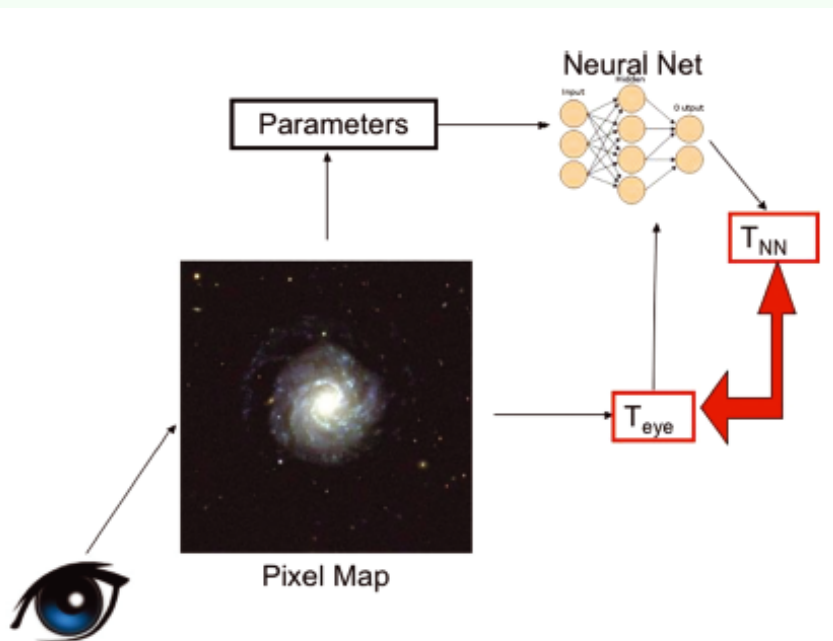
Coming to a  
bookstore near you  
in February 2012

<http://www.giss.nasa.gov/staff/mway/book/>

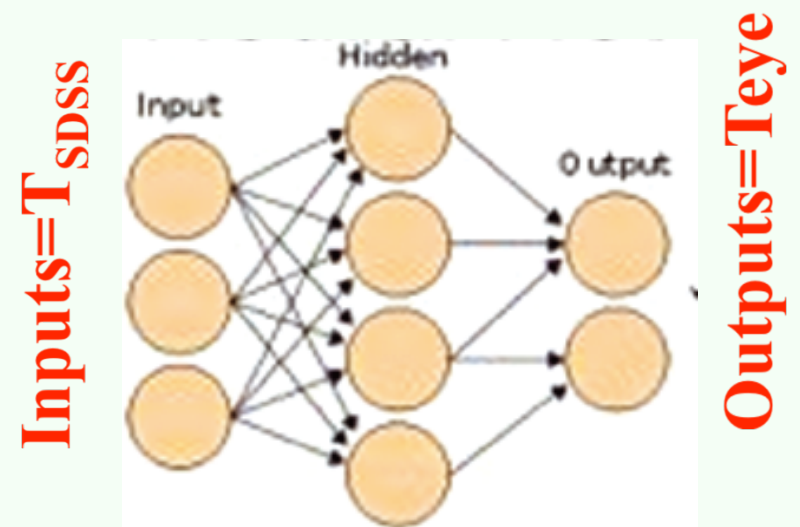
# Crowd-Sourcing + Machine Learning

**“Galaxy Zoo: reproducing morphologies via machine learning”, Banerji et al. 2010**

1. GalaxyZoo morphologies from volunteers ( $T_{\text{eye}}$ )
2. Primary & Secondary Isophotal Parameters ( $T_{\text{SDSS}}$ )



**Neural network**



# Primary & Secondary Isophotal Parameters ( $T_{\text{SDSS}}$ )

**Table 1.** First set of input parameters based on colours and profile fitting.

Name	Description
<i>dered_g-dered_r</i>	$(g - r)$ colour
<i>dered_r-dered_i</i>	$(r - i)$ colour
<i>deVAB_i</i>	de Vaucouleurs fit axial ratio
<i>expAB_i</i>	Exponential fit axial ratio
<i>lnLexp_i</i>	Exponential disc fit log likelihood
<i>lnLdeV_i</i>	de Vaucouleurs fit log likelihood
<i>lnLstar_i</i>	Star log likelihood

**Table 2.** Second set of input parameters based on adaptive moments.

Name	Description
<i>petroR90_i/petroR50_i</i>	Concentration
<i>mRrCc_i</i>	Adaptive (+) shape measure
<i>aE_i</i>	Adaptive ellipticity
<i>mCr4_i</i>	Adaptive fourth moment
<i>texture_i</i>	Texture parameter

**Table 5.** Summary of results for the entire sample when using input parameters specified in Tables 1 and 2.

		Galaxy Zoo		
		Early type (per cent)	Spiral (per cent)	Point source/artefact (per cent)
A	Early type	92	0.07	0.6
N	Spiral	0.1	92	0.08
N	Point source/artefact	0.2	0.2	96

Results!

# Cultural Changes

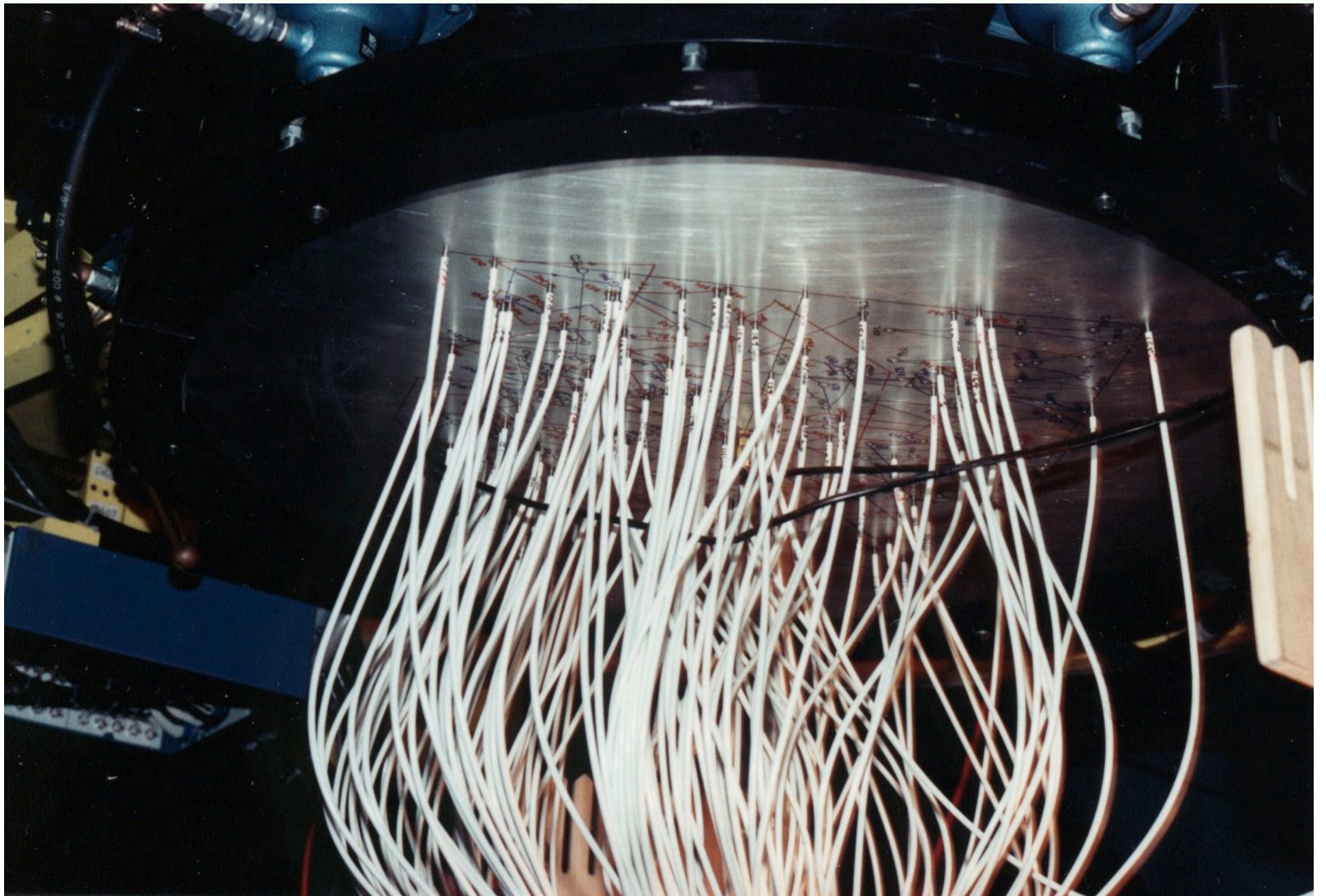
## How are these things changing the way we work?

- Avoid applying for grants \*and\* telescope time and requisite travel funding
- Avoid specialized instrument knowledge, data acquisition, data reduction, backup/storage concerns
  - Positive and Negative aspects
- More time for thinking up good questions?
  - Submit a query (10 min)
  - Download the data (2 min)
  - Write your paper

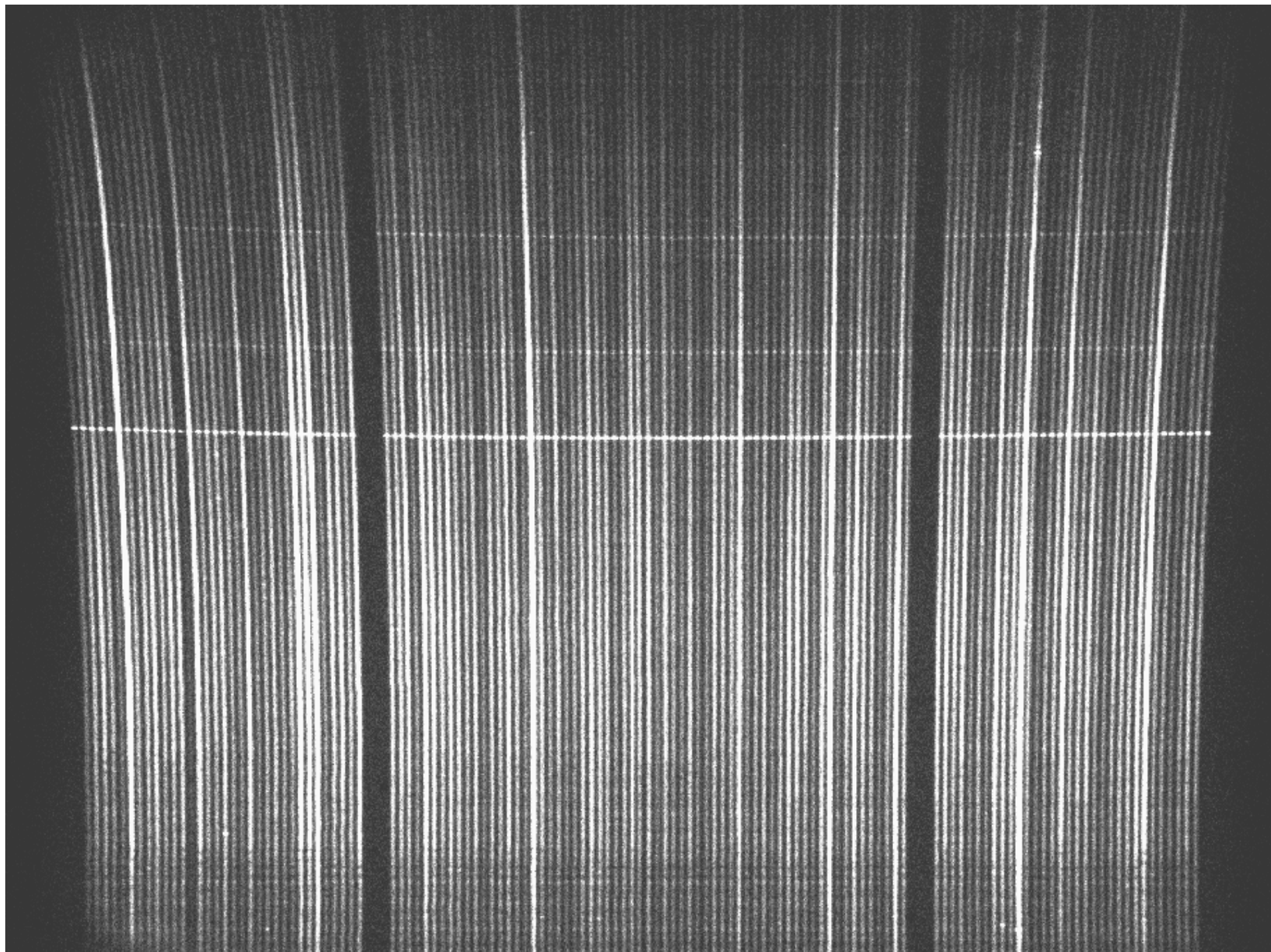


# My world... then

- PhD: Spent 2 years in Chile collecting spectroscopic redshifts in (Abell) clusters of galaxies
  - Incredibly specialized instrument knowledge
    - Detector was photon counting, but also imaging like CCD
    - Fiber spectrograph: instrument calibration, sky subtraction, etc.
  - Heavy use of time/labor
    - Plates drilled in Pasadena, shipped to Chile weeks before
    - A 3 night run required showing up 3-4 days before to prepare
      - Marking up plates, checking out instrument, getting to observatory
    - *\*very\** long nights: plugging plates beforehand, collecting calibration frames (sky flats) until well after twilight
  - 1—2 years “reducing” the data for thesis







# My world... today

“Galaxy Zoo Morphology & Photometric Redshifts in the Sloan Digital Sky Survey”

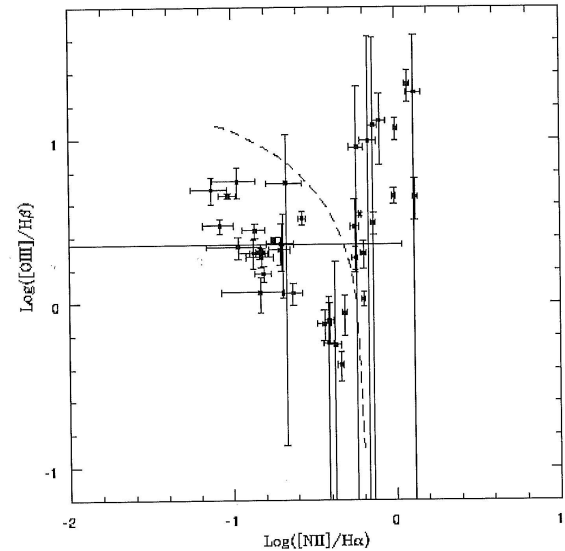
- 1) SDSS Data Release 7 (Oct 2008)
- 2) GalaxyZoo Data Release 1 (Feb 2011)
  - Morphologies for SDSS galaxies
  - Banerji et al. 2010 isophotal parameters of use
- 3) Gaussian Process Regression (Foster et al. 2009)
- 4) Cross-match catalog built in 5 min (March 2011)
- 5) Paper written in 2 weeks
  - Received March 25, 2011 : Accepted April 21, 2011



# Differences in my data compression?

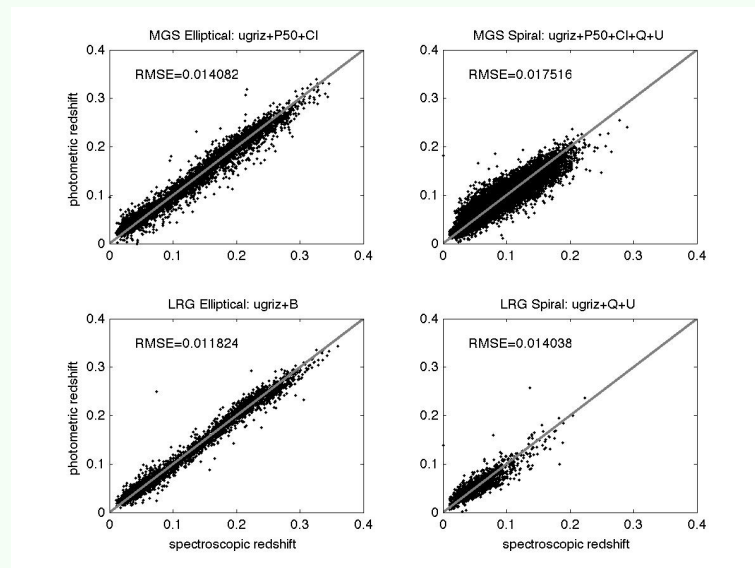
Thesis:

40 DAT (~100GB) + 4 years →



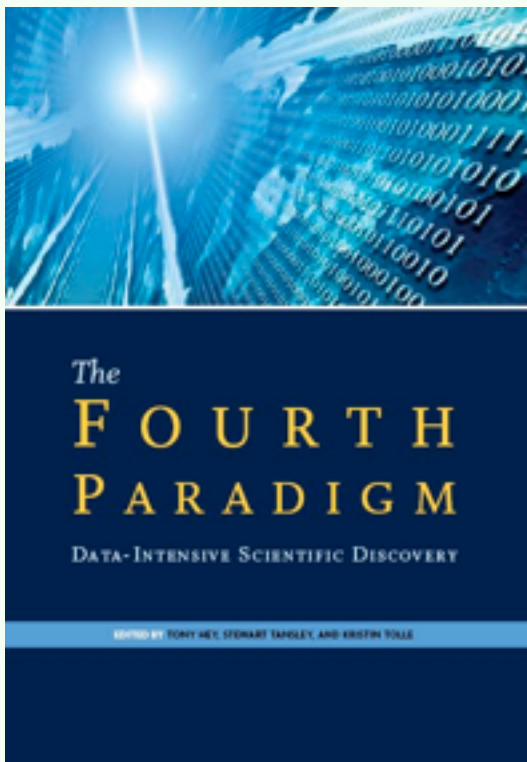
Last Project:

30TB + 1000 years? →



# I'm not a Chauvinist!

These changes/challenges are not limited to Astronomy



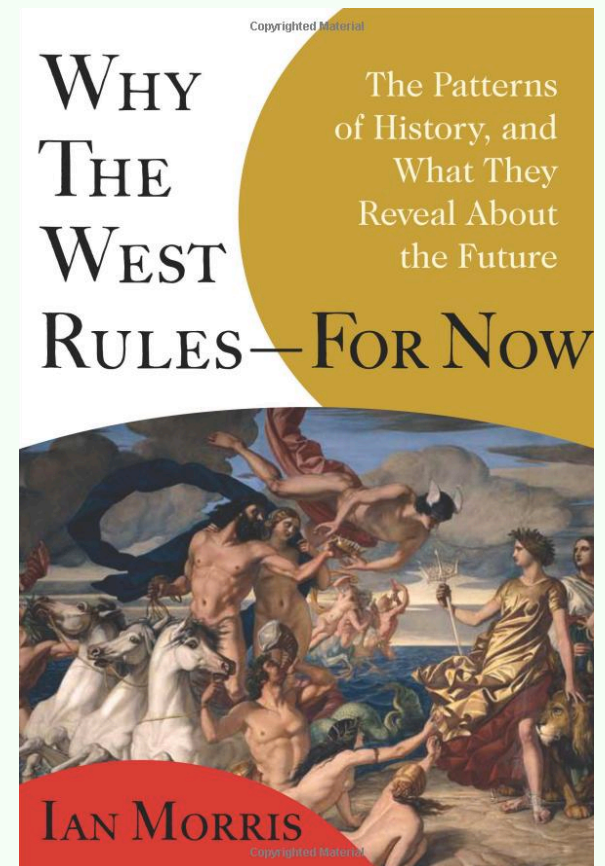
<http://www.sciencemag.org/site/special/data>

# The Humanities

**Nor are these changes limited to Science!!**

We know **data** is also allowing new collaborations within the humanities & even between science and the humanities:

- History
- Sociology
- Anthropology, Archaeology -- Carbon Dating
- Geology
- Genetics



# Nor are these changes going to solve all of our problems...

## Six Provocations for Big Data

[danah boyd](#)

Microsoft Research; University of New South Wales (UNSW); Harvard University - ~~Berkman~~ Center for Internet & Society

[Kate Crawford](#)

University of New South Wales (UNSW)

September 21, 2011

*A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society*

### Abstract:

The era of Big Data has begun. Computer scientists, physicists, economists, mathematicians, political scientists, bio-informaticists, sociologists, and many others are clamoring for access to the massive quantities of information produced by and about people, things, and their interactions. Diverse groups argue about the potential benefits and costs of analyzing information from Twitter, Google, Verizon, 23andMe, Facebook, Wikipedia, and every space where large groups of people leave digital traces and deposit data. Significant questions emerge. Will large-scale analysis of DNA help cure diseases? Or will it usher in a new wave of medical inequality? Will data analytics help make people's access to information more efficient and effective? Or will it be used to track protesters in the streets of major cities? Will it transform how we study human communication and culture, or narrow the palette of research options and alter what 'research' means? Some or all of the above?

This essay offers six provocations that we hope can spark conversations about the issues of Big Data. Given the rise of Big Data as both a phenomenon and a methodological persuasion, we believe that it is time to start critically interrogating this phenomenon, its assumptions, and its biases.

(This paper was presented at Oxford Internet Institute's "A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society" on September 21, 2011.)



We need to open a discourse – where there is no effective discourse now – about the varying temporalities, spatialities and materialities that we might represent in our databases, with a view to designing for maximum flexibility and allowing as possible for an emergent polyphony and polychrony. *Raw data is both an oxymoron and a bad idea; to the contrary, data should be cooked with care.*

**Geoffrey Bowker (2005, p. 183-184)**

# Conclusions

1. Science is becoming more data intensive
2. We are exploring new collaborations to deal with the data deluge forced upon us by technological advances
3. This also leads to rethinking our methodologies:
  - Probability theory, statistics, machine learning, data mining, etc...
  - And perhaps our productivity?